Sémantika přirozeného jazyka a reálného světa – počítačové zpracování

Roman Mouček, Pavel Mautner

Abstrakt: *Článek se zabývá možnostmi počítačového zpracování sémantiky přirozeného jazyka a reálného světa a pokládá otázku, do jaké míry je toto zpracování možné a smysluplné. Odpověď pak hledá v kombinaci poznatků a zkušeností tří různých oborů – neurověd, lingvistiky a informatiky. Stručně je prezentován pohled neurověd na fungování lidského mozku a popsány paměťových složky mající vliv na zpracování sémantiky přirozeného jazyka a sémantiky vnějšího reálného světa. Krátce je představen i vnější, lingvistický pohled na přirozený jazyk a jeho sémantické roviny. Z informatických oborů jsou pak představeny přístupy umělé inteligence a softwarového inženýrství. Zmíněna je i vize sémantického webu.*

Abstract: *This article deals with possibilities of computerized processing of natural language and real world semantics. The meaningful level of this processing is looked for by combining knowledge and experience from three various scientific fields – neuroscience, linguistics and informatics. The view of neuroscience on the functioning of human brain is presented and brain memories having influence on processing of natural language semantics and outer real world are shortly described. Also the view of linguistics on natural language and its semantic layers is briefly introduced. Informatics is represented by the fields of artificial intelligence and software engineering. The idea of semantic web is finally mentioned.*

Úvod

Zpracování sémantiky přirozeného jazyka[1] patří mezi problémy, se kterými si současné počítačové systémy a aplikace dokážou poradit jen částečně. Teorií, přístupů i experimentů, které se pokoušejí přirozený jazyk popsat a zpracovat i na sémantické úrovni, existuje samozřejmě velké množství, málokdy se však ptáme, proč se o vyřešení tohoto problému vůbec snažíme. Zvykli jsme si zadávat klíčová slova do internetových vyhledavačů, při telefonátech na linky mobilních operátorů či dopravních společností hledáme často nejkratší cestu k živému operátorovi, abychom se nemuseli „bavit“ s počítačovým dialogovým systémem, a jestli jsme nalezli ten správný dokument, který potřebujeme, poznáme nejlépe sami po alespoň letmém přečtení. Navíc si často nerozumíme ani mezi sebou (nedokážeme správně interpretovat promluvu jiného člověka). Existují kromě touhy po počítačovém „pokoření“ přirozeného jazyka i racionální důvody, proč se snažíme počítačově zpracovat i jeho sémantiku? Co nám toto zpracování může přinést a do jaké míry jej máme realizovat?

Sémantiku přirozeného jazyka můžeme chápat jako jazykovou interpretaci složitého vnějšího reálného světa a zároveň složitého vnitřního světa jednotlivce. Kromě smyslového zpracování vnější reality se na výsledné jazykové interpretaci podílí i řada interních procesů (schopnost kognitivního a emociálního zpracování, zkušenost, stupeň zvládnutí daného jazyka apod.). Výsledné lidské zpracování sémantiky přirozeného jazyka je stejně jako další aspekty lidského chování a jednání ovlivněno jak geneticky, tak učením a životní zkušeností.

[1] Počítačové zpracování sémantiky přirozeného jazyka tak, jak je chápané v tomto textu, zahrnuje schopnost počítačového systému smysluplně interpretovat text či promluvu v přirozeném jazyce a posléze na tento text či promluvu adekvátně reagovat; zkoumána je tedy schopnost počítačového „porozumění“ přirozenému jazyku.

Historicky lze počítačové zpracování sémantiky přirozeného jazyka spojit s vývojem řečových aplikací a automatických překladačů[2]. V aplikacích umělé inteligence je sémantická úroveň jazyka řešena např. při vývoji dialogových informačních systémů založených na rozpoznávání přirozeného mluveného jazyka. Tyto systémy a aplikace se však v případě mluveného jazyka zaměřují většinou na zpracování signálu, určení slovních hypotéz a částečné zpracování syntaxe promluvy. Sémantika, pokud je vůbec zpracovávána, se obvykle omezuje na vyhledávání a porovnání významu slova vzhledem k systémové či doménové databázi. Systémy aplikující složité systémy pravidel, speciální formalismy, historii dialogu, pravděpodobnostní modely promluv apod. se objevily pouze v experimentální podobě (podrobněji např. [5]). Úspěšnější z existujících systémů pak „paradoxně" využívají ke zpracování sémantiky hrubší sílu (rozsáhlé korpusy, statistika, jednoduchá pravidla) nežli sofistikované algoritmy a systémy složitých pravidel.

Dnes je pozornost věnována především počítačovém zpracování psaných dokumentů – vyhledávání informací v dokumentech, metodám organizace a klasifikace dokumentů (např. projekt WEBSOM [2]), vytváření významových sítí (např. projekt EuroWordNet [2]), anotaci korpusů, sumarizaci dokumentů, automatickému překladu apod. O zpracování sémantiky ve smyslu porozumění však často nejde. Řada metod přistupuje k dokumentu jako k souboru dat, který je nutné porovnat s jiným souborem dat (např. s dotazem uživatele nebo s jiným dokumentem), obsah souboru a jeho význam však dále nijak nezkoumá.

Velkou výzvou a potencionálním řešením problémů pří pokusech o zpracování sémantiky přirozeného jazyka se stala snaha o realizaci tzv. sémantického webu (sekce 3). Zde však již nemůžeme hovořit o přirozeném jazyku a jeho zpracování (centrem zpracování není dokument), ale o datovém modelování reálného světa (zpracovávána jsou označkovaná data) a práce s konceptualizovanými daty ve webovém prostředí. I tato myšlenka a především pokus o realizaci (zaštiťovaný konsorciem W3C) však zatím nepřinesla očekávané výsledky.

„Popovídat si" s počítačem v přirozeném jazyce (ať textově nebo hlasově) je dnes stále možné jen ve velmi omezené doméně. Přes snahu odborníků z různých vědních disciplín navrhnout formalismy či metajazyky popisující libovolnou část sémantické roviny jazyka, nedošlo v tomto směru k významnějšímu pokroku. Proto je i v současné době vývoj aplikací, které zpracovávají sémantickou informaci přirozeného jazyka, velmi problematický. Proč je přirozený jazyk takto nezkrotný? Lze se v této situaci poučit u člověka, aneb zkoumat, jakým způsobem zpracovává sémantiku lidský mozek? Je možné, že plné zpracování sémantiky přirozeného jazyka (jakéhokoli dokumentu či promluvy) současnými prostředky výpočetní techniky je nezvládnutelný úkol?

1 Sémantika a lidský mozek

Z poznatků neurověd vyplývá, že funkční systémy lidského mozku (např. systém jazyk a řeč nebo paměť) vykazují znaky rozsáhlého distribuovaného systému, který je funkčně částečně modulární (podrobněji např. [4]). Tyto funkční systémy mají tzv. zúžené profily informačního chodu[3]. Součástí dlouhodobé vědomé paměti jsou i tzv. epizodická a sémantická paměť. Epizodická paměť je paměťová složka sloužící k zapamatování událostí vázaných na kontext, prostor a čas (události autobiografického charakteru), sémantická paměť slouží

[2] V tomto případě je zpracovávána jak sémantika reálného světa, který se snaží přirozený jazyk postihnout (dialogový systém – modelování domény), tak sémantika samotného jazyka (dialogový systém – porozumění promluvě, automatický překlad).

[3] Zúžený profil informačního chodu označuje místo, jehož poškození je kritické vzhledem k funkci příslušného funkčního systému.

k zapamatování faktů, pojmů a významů „kontextově nezávislých“. Obě paměti jsou částečně vázány na odlišné části mozku a částečně se překrývají. Vykazují vzájemnou spolupráci, jsou však schopné fungovat i samostatně např. při významném poškození jedné z nich. Kromě epizodické a sémantické paměti lze identifikovat i další částečně oddělené paměťové složky, které se podílejí na zpracování přirozeného jazyka, např. lexikální a syntaktické informace[4]. Na výslednou interpretaci sémantiky přirozeného jazyka má vliv i emoční paměť.

Jestliže budeme předpokládat, že epizodická i sémantická složka paměti mají vliv na výsledné promluvy jednotlivce i na dokumenty, které tento jednotlivec napíše, pak stejně tak mají tyto paměťové složky vliv na interpretaci promluv jiných osob či dokumentů napsaných jinými lidmi. Jestliže vliv sémantické paměťové složky skýtá možnost smysluplného zpracování výsledných promluv nebo dokumentů, pak příspěvek epizodické složky paměti je pro počítačové zpracování sémantiky patrně neřešitelným problémem. Sepětí této paměťové složky s osobní zkušeností jednotlivce, která se navíc neustále vyvíjí a mění v čase, by znamenalo nutnost přizpůsobení počítačového systému této zkušenosti. Umíme dostatečně přesně modelovat všechny aspekty zkušenosti? Jsme schopni v reálném čase tyto údaje počítači předávat?

Větší možnosti nám poskytuje sémantická složka paměti. Nezávislost na kontextu odstiňuje do značné míry osobní zkušenost, přesto by bylo iluzorní obsah sémantické složky paměti považovat za shodný u všech lidských jedinců. Můžeme však předpokládat, že vzájemná shoda bude narůstat u lidí žijících v stejném časovém období, ve stejné kulturní oblasti, v podobných sociálních podmínkách, s podobnou úrovní vzdělání atd. Vzájemnou shodu sémantické složky paměti podporuje v některých životních sférách i postupující globalizace.

Položky sémantické složky paměti vykazují vzájemný asociativní vztah na obecnější rovině. Tento vztah založený na „chaotickém“ propojení obrovského množství synapsí je v reálných aplikacích modelován mnoha prostředky a formalismy (pravidla, gramatiky, umělé neuronové sítě, rámce, sémantické sítě, objektově orientovaný návrh apod.). Avšak každý z těchto prostředků se osvědčil pouze ve velmi specifických úlohách. Při experimentech na rozsáhlejších doménách dochází k nekontrolovatelnému nárůstu reálně nesmyslných vazeb.

Jak je možné, že v sémantické složce paměti vznikají „nesmyslné“ vazby v menší míře než při použití modelovacích prostředků? Můžeme předpokládat, že vyšší reálnost sémantické složky paměti je výsledkem mohutné výpočetní kapacity neuronálních obvodů? Znamená to, že modelovací prostředky jsou nepřesné jen proto, že obsahují málo výpočetních prvků? Znamená to, že i základem lidského usuzování není nic jiného než dostatečný počet neuronů a synaptických spojení vznikajících nejprve na genetickém základě a poté na základě mapování reálných událostí nejprve do epizodické a posléze sémantické složky paměti? Znamená to, že pokud nemáme k dispozici dostatečnou výpočetní kapacitu, pak nemá smysl se pokoušet vymýšlet sofistikované formalismy pro zpracování přirozeného jazyka, neboť ty pak nezvládnou rozsah uchovávaných znalostí, nebo naopak produkují množství nesmyslných znalostí? Statistické metody jsou např. ve srovnání s formalismy založenými na systému pravidel relativně úspěšné, jejich spolehlivost roste s množstvím dostupných trénovacích dat.

2 Sémantika a lingvistika

Lingvistické teorie obecně zkoumají přirozený jazyk na základě jeho vnějších projevů. Sémantika jako lingvistická disciplína zaujímá ke zpracování sémantické informace

[4] Některé experimenty připouštějí existenci hierarchie objektů ve spánkových lalocích.

přirozeného jazyka rozdílné postoje. Připouští, že vágnost a nejednoznačnost přirozeného jazyka je jeho základní vlastnost (tato teorie odpovídá konceptu paměťových složek, sekce 1). Tento přístup pak kontrastuje se snahou lingvistiky vybudovat obecnější úroveň reprezentace sémantiky a obecnou sémantickou hierarchii. Někteří lingvisté se pak pokoušejí najít kontextově nezávislou úroveň přirozeného jazyka (odpovídá „globalizované“ sémantické složce paměti), zatímco jiní přiznávají mnohoznačnost tzv. jednoduchého jazykové znaku[5], závislost významu na doméně (odpovídá sémantické složce paměti v rámci určité domény), situaci a individuální interpretaci (odpovídá epizodické složce paměti). Pak jen tzv. konceptuální a kolokační významová složka (vysvětleno např. v [5]) jednoduchého jazykového znaku může sloužit jako základ pro vytváření sémantických hierarchií a ontologií.

Sémantické teorie složitého jazykového znaku (promluvy) jsou velmi různorodé a neucelené. Připouštějí (podobně jako v případě jednoduchého jazykové znaku) jak existenci a neoddělitelnost jednotlivých významových složek, tak možnost pracovat s několika částečně nezávislými stupni porozumění. Výsledné popisy přirozeného jazyka a jeho sémantických rysů bývají buď vágní, a tedy obtížně aplikovatelné pro počítačové zpracování přirozeného jazyka, nebo obsahují značné množství pravidel a výjimek, a počítačové zpracování pak generuje značné množství nereálných jazykových konstrukcí či jejich interpretací (problém vzájemného mapování abstraktních vrstev popisu sémantiky, aplikace gramatik ad.). Mezi teorie zabývající se sémantikou přirozeného jazyka patří např. kontextový přístup, princip kompozicionality, syntakticko-sémantické větné vzorce a mimojazykové mikrosituace (podrobně v [3]), model lidského chování nebo teorie aktuálního členění věty v češtině (podrobně v [8]).

Moderní lingvistika se zabývá i proměnlivostí sémantiky přirozeného jazyka a sémantického pole v čase, tedy posunu interpretace významu promluv a textů u skupiny lidí, která jazyk používá v různých časových obdobích.

3 Sémantika a informatika

Informatika se od svého vzniku věnuje především jazykům formálním. Přesto prudký rozvoj tohoto oboru vedl až k situaci, kdy jazykem formálním chceme zpracovávat jazyk přirozený. Zpracováním sémantiky přirozeného jazyka se v rámci informatiky nejprve zabývala umělá inteligence. Sémantické reprezentační a interpretační systémy vytvořené v rámci vývoje inteligentních dialogových systémů (přehled např. [5]) se vyznačují mnohými společnými vlastnostmi. Významné je především centrální postavení slovesa a volba abstraktní úrovně popisu významu ostatních složek věty či promluvy (nazývané koncepty, tematické role,…). Jelikož užití pouze jedné abstraktní úrovně popisu významu se ukazovalo jako nedostatečné, tvůrci dialogových informačních systémů přistupovali k definici několika abstraktních úrovní jak doménově závislých, tak doménově nezávislých. Vyvíjené metody umělé inteligence pro (nejen) zpracování sémantiky jsou dobře použitelné v úzkých a specifických situacích, v případě zpracování rozsáhlejší domény se i v této oblasti stále více prosazují metody založené na statistice, rozsáhlé datové základně a výkonném hardwaru.

Významný krok vpřed při modelování sémantiky reálného světa (a některých významových složek přirozeného jazyka) udělalo i softwarové inženýrství. Nelze zde samozřejmě hovořit o zpracování sémantiky přirozeného jazyka, lze však pozorovat zvyšující se úroveň vyjádřitelné abstrakce, kterou dovoluje využít zvolený formální jazyk. Prostředky objektově orientované analýzy a objektově orientovaného návrhu tak poskytují aparát pro částečný popis

[5] Pojmy jazykový znak, jednoduchý a složitý jazykový znak vysvětleny podrobně např. v [5].

sémantické paměťové složky (typicky asociativní vazba). Tento popis je samozřejmě zjednodušený, přesto však dostupnější a srozumitelnější širší odborné komunitě nežli prostředky a metody umělé inteligence.

Objektově orientovaná analýza definuje pohled na reálný svět bez ohledu na implementační prostředky. Analytický model vystihuje podstatu omezeného světa ve formě tříd objektů a jejich vzájemných asociativních vazeb. Je možné jej přirovnat k abstrakci sémantické složky paměti definující základní koncepty (pokud přijmeme i lexikální složku paměti, pak je modelována i tato) a jejich vzájemné asociace. Objektově orientovaný návrh pak analytický model dále rozšiřuje o definici vlastností a chování typizovaných objektů a o vzájemnou interakci těchto objektů. Modelování vlastností se zde liší od prostředků umělé inteligence. Zatímco v klasických formalismech jsou vlastnosti často definovány na stejné úrovni abstrakce jako související objekty, zde se stávají přímo součástí objektového paradigmatu a principu zapouzdření. Dalším modelovacím prostředkem je pak princip dědičnosti typů objektů (zavedení hierarchie).

Nalezneme podobné struktury i v oblastech sémantické složky paměti? Lze aktivaci neuronových oblastí v daném čase úspěšně modelovat definicí vlastností a způsobů chování typizovaných objektů? Můžeme předpokládat, že dva jevy prezentované na neuronální úrovni aktivací částečně společných souborů neuronů budou na strukturální úrovni prezentované alespoň částečně stejnými vlastnostmi a chováním[6]? Z hlediska modelování reálného světa výše popsaným strukturálním způsobem je zřejmá snaha související jevy udržet pohromadě ať asociativními vazbami, či přímo na úrovni definice vlastností a chování objektů.

Další pohled na softwarovou realizaci jednotlivých složek paměti představuje model uchování persistentních objektů. Sémantika persistentních dat je nejčastěji definována relačním modelem (v případě použití databáze) či tzv. sémantickými značkami na úrovni XML dokumentu. Z hlediska dalšího zpracování dat se však volnost při definici sémantických značek ukazuje jako nepraktická, neboť je nutná jejich další interpretace. Obecně pak hrozí nebezpečí vzniku několika vrstev popisů - metadat na úrovni různých vrstev abstrakce. Mapování mezi jednotlivými vrstvami abstrakce je pak obecně velmi problematické.

Za významný mezistupeň (a možný bod setkání akademičtější umělé inteligence a praktičtějšího softwarového inženýrství) mezi komplexnějším zpracováním sémantiky přirozeného jazyka a typickým modelováním sémantiky reálného světa počítačovými systémy lze považovat vizi tzv. sémantického webu. Sémantický web [7] lze charakterizovat jako rozšíření současného webu takovým způsobem, že bude možná kombinace a integrace dat z různých zdrojů, a tak se významně zlepší spolupráce jak mezi lidmi, tak mezi počítačovými systémy. Tento přístup zahrnuje myšlenkový posun od zpracování a výměny dokumentů ke zpracování a výměně dat. Tato vize předpokládá, že data prezentovaná na internetu budou mít přesně definovaný význam, a tím bude umožněno jejich strojové zpracování. Celá myšlenka sémantického webu tak počítá s konceptualizací dat (existencí doménových ontologií), existencí aktivních inteligentních komponent zabezpečujících požadavky uživatelů a standardizovaného popisu webových zdrojů[7]. Idea sémantického webu byla představena již v roce 2001 [1] a je podporována konsorciem W3C.

[6] Experimenty potvrzují, že související jevy aktivují částečně společné skupiny neuronů.

[7] Práce s webem by byla obdobná práci s relační databází; význačným přínosem je značná relevance odpovědi na položený dotaz.

4 Závěr

Současné přístupy k počítačovému zpracování sémantiky reálného světa a přirozeného jazyka se střetávají s mnoha problémy a otázka možností zpracování sémantiky na úrovni strojového porozumění je stále otevřená. Přesto je z kombinace poznatků a zkušeností různých oborů (neurověd, lingvistika, informatika) zřejmé, kde se nachází v současné době realizovatelná hranice. Modelování světa a přirozeného jazyka jsou doménou oborů umělé inteligence a softwarového inženýrství. Zatímco klasická, abstraktněji orientovaná umělá inteligence staví na formalismech a zobecnění na úrovni gramatik či různých systémů pravidel, softwarové inženýrství využívá modelovacích prostředků prakticky využitelných při realizaci konkrétních aplikací ve velmi omezených doménách. Oba obory se pak potkávají při využití statistických metod či hardwarového výkonu. Jejich vzájemné provázání je pak pravděpodobné i při realizaci myšlenek sémantického webu.

Vezmeme-li v úvahu např. uspořádání neuronů a synapsí v lidském mozku, existenci asociativních vazeb, spolupráci sémantické a epizodické paměti, zásadní roli epizodické paměti v běžných životních situacích, emoční paměť jako významný interpretační mechanismus nebo zkušenosti s modelováním reálného světa v počítačových systémech, je jen velmi obtížné si představit komplexní zpracování přirozeného jazyka a vytvoření obecné abstraktní úrovně sémantické reprezentace přirozeného jazyka. I v případě modelování sémantiky jednoduchých domén jsou vytvořené modely velmi složité a neodrážejí komplexitu celé situace, či zavádějí nereálné vazby, vlastnosti a způsoby chování. Je tedy možné, že jediným modelovacím prostředkem pro popis a porozumění přirozenému jazyku je prostředek výpočetní kapacitou, uspořádáním a fungováním obdobný lidskému mozku.

Na začátku jsme se ptali, proč se snažíme počítačově zpracovávat přirozený jazyk. Možná to nejen neumíme, ale ani nepotřebujeme, a např. vize sémantického webu může určovat hranici, ke které z pohledu zpracování sémantiky jazyka a reálného světa stačí v informatice dojít.

Poděkování:

Tato práce vznikla v rámci řešení projektu MŠMT č. 2C06009 „Prostředky tvorby komplexní báze znalostí pro komunikaci se sémantickým webem v přirozeném jazyce“.

Literatura:

[1] Berners-Lee T., Hendler J., Lassila O.: *The Semantic Web,* Scientific American, May 2001, dostupné online http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2 (8.8. 2008)
[2] EuroWordNet, dostupné online http://www.illc.uva.nl/EuroWordNet/ (13. 8. 2008)
[3] Grepl M., Karlík P.: *Skladba češtiny*, Olomouc, 1998
[4] Koukolík F.: *Paměť a její poruchy*, *Lidský mozek. Funkční systémy. Norma a poruchy*, Portál, Praha, 2002
[5] Machová S., Švehlová M.: *Sémantika & Pragmatická lingvistika*, Univerzita Karlova, Praha, 2001
[6] Mouček R.: *Sémantika v dialogových systémech,* disertační práce, Západočeská univerzita v Plzni, Plzeň, 2004
[7] Semantic Web, dostupné online http://www.w3.org/2001/sw (8.8. 2008)

[8] Sgall P., Hajičová E., Panevová J.: *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, Academia, Praha, 1986
[9] WEBSOM, dostupné online http://websom.hut.fi/websom/ (13. 8. 2008)

Adresa:

Ing. Roman Mouček, Ph.D.
Katedra informatiky a výpočetní techniky, Fakulta aplikovaných věd, Západočeská univerzita v Plzni,
Univerzitní 8, 306 14 Plzeň
E – mail: moucek@kiv.zcu.cz

Ing. Pavel Mautner, Ph.D.
Katedra informatiky a výpočetní techniky, Fakulta aplikovaných věd, Západočeská univerzita v Plzni,
Univerzitní 8, 306 14 Plzeň
E – mail: mautner@kiv.zcu.cz